Statistical methods for epigenomic data:
Studying the importance of
3D chromatin structure and DNA-methylation.

by Tonje Gulbrandsen Lien

Dissertation presented for the degree of
Philosophiae Doctor (PhD)

Department of Mathematics
University of Oslo
2015

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1679

ISSN 1501-7710

Cover: Hanne Baadsgaard Utigard.
Print production: John Grieg AS, Bergen.

Produced in co-operation with Akademika Publishing.
The thesis is produced by Akademika Publishing merely in connection with the thesis defence. Kindly direct all inquiries regarding the thesis to the copyright holder or the unit which grants the doctorate.

Acknowledgements

When working with analysis of epigenomic data, a range of different research fields is combined. The need for expert knowledge in medicine, biology, informatics and statistics is clear. I am fortunate to have had the opportunity to work with researchers from all these different fields. It is challenging and crucial that these different research areas are combined and speak each other's "language". I find these challenges exciting and intriguing and I am looking forward to further collaboration.

I would like to give a special thank to Jonas Paulsen, and many more at the Department of Informatics. You welcomed me in my first year and I really had a good time working with you on Paper I and other projects. In my last part of my third year, I was particularly happy to get the opportunity to visit Mark van de Wiel, professor in statistics for genomics at Vrije Universiteit in Amsterdam. You have been an excellent extra supervisor. I also appreciated the many productive and inviting discussions with professor Kaare M. Gautvik and senior researcher Sjur Reppe. There have been other projects that unfortunately did not end in any papers, but I still feel I gained a lot from those collaborations. Lastly, I am very grateful to have had both professor Ingrid K. Glad and professor Ørnulf Borgan as supervisors. I always looked forward to our meetings and I have learned so much from you.

In addition, a warm thank to my friend, colleague and now post doc Kristoffer Hellton, who I have had many important lunches with where we discussed all types of challenges coming our way. At home, my family have always been there for me and given me all the extra support I needed. And Erlend, thank you for being in my life, for being patient and for being the wonderful, magnificent person you are!

Blindern, July 2015
Tonje G. Lien

List of papers

Paper I

Jonas Paulsen*, Tonje G. Lien*, Geir Kjetil Sandve, Lars Holden, Ørnulf Borgan, Ingrid K. Glad and Eivind Hovig (2013). Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic Acids Research*, 41(10): pp 5164–5174. *Joint first authors.

Paper II

Sjur Reppe, Tonje G. Lien, Vigdis T. Gautvik, Ole K. Olstad, Hege G. Bakke, Robert Lyle, Marianne Kringen, Ingrid K. Glad and Kaare M. Gautvik (2015). DNA methylations in bone correlate markedly to BMD associated transcript levels and distinguish between osteoporotic and healthy postmenopausal women *(Manuscript)*.

Paper III

Mark A. van de Wiel, Tonje G. Lien, Wina Verlaat, Wessel N. van Wieringen and Saskia M. Wilting (2015). Better prediction by use of co-data: Adaptive group-regularized ridge regression. *Statistics in Medicine (Published)*.

Paper IV

Tonje G. Lien, Sjur Reppe, Kaare M. Gautvik, Ørnulf Borgan and Ingrid K. Glad (2015). Integration of epigenomic and genomic data in high-dimensional penalized regression. A cohort study on bone mineral density *(Manuscript)*.

Contents

1 Introduction

Statistics is the science of analyzing and interpreting data. Extensive amounts of data are being collected and stored from various areas; from example finance and climate to genomics. Typically, with the rapid technology development data becomes more complex, with large number of variables and few samples, strong correlations, complex interactions between different data sets and higher signal-to-noise ratio.

In biostatistics, epigenomics is a field of growing interest with complex data structures. Epigenomics is the science of what our genes are "wearing"; the elements surrounding the gene which can change gene expression and biological cell functions (Quina et al., 2006). The entire complex consisting of DNA and the molecules attached to it, such as proteins and RNA, is called chromatin. Chromatin is tightly packed within the nucleus in each cell in a specific manner and its spatial structure is of major importance for biological processes. For instance, Kleinjan and van Heyningen (2005) and West and Fraser (2005) showed that three-dimensional (3D) contacts between different parts of DNA are associated with gene expression. Sophisticated statistical methods are needed to analyze the spatial structure of chromatin and questions such as if two points on the genome are more co-localized than expected by chance, can be answered using a well-defined hypothesis tests. The complex distribution of the spatial structure poses a challenge since the distribution of the 3D structure is far from homogeneous and specific contacts are highly correlated. The difficulty of finding a proper distribution for the data calls for permutation testing and randomization procedures.

Another important epigenetic factor is the simple methylation group which can get attached to the DNA strands. DNA-methylation strongly influences the gene expression, and when the start site or promoter region of a gene is methylated, the gene is typically not expressed. Therefore, DNA-methylations have an enormous potential to influence biological function in the cells and human body. When analyzing the effect of genome wide DNA-methylations on an outcome, there are typically more variables than samples. In this high dimensional setting, penalized regression is a much used statistical tool. But, with a large amount of variables, and possibly a great level of noise in the data, the penalized regression can benefit from some guidance, for instance in the form of co-data, such as gene expressions. But how to find the optimal way to integrate and analyze complex and large data sources is a great challenge.

The outline of this thesis is as follows: Section 2 presents an introduction to genomics and epigomics and questions related to the analyses of chromatin structure and DNA-methylation. Section 3 presents a short overview of the main aims of the thesis. The background methodology is presented in Section 4, introducing hypothesis testing and penalized regression. Section 5 gives a summary of the four papers, and Section 6 some extensions and future work. Lastly, a

discussion of relevant themes in the thesis are presented in Section 7.

2 Biological background

DNA is the genomic material, unique to all humans (except identical twins), containing our inheritable potential. DNA consists of nucleotides, which are complex molecules of sugar (deoxyribose), phosphate and one of the four bases, cytosine (C), thymine (T), adenine (A) and guanine (G). DNA contains 3.3 billion nucleotides (Ziegler et al., 2010). With a specific linear ordering, the nucleotides form coding sequences. In 2003, the first complete sequencing of a human genome was finished by the Human genome project (Collins et al., 2003; Watson, 1990), presenting a typical reference human genome. With this event, the "book of life" had been read, but it was only the beginning and the next step has been to understand what type of influence the coding could have. The linear ordering of all nucleotides are identical in every single cell in an organism, except for red blood cells (Ziegler et al., 2010). But, even with identical DNA, there is a great variety of cell types with different tasks and functions. In the last decades, there have been done tremendous work to find the explanation of this differently functioning cells.

2.1 Genomics and epigenomics

The term genomic is used for the study of genes, genetic variation and their heredity in living organisms and focuses on the DNA sequence. There are between 20000 and 25000 genes in the human DNA (Pertea and Salzberg, 2010). The most common definition of a gene is a sequence of nucleotides providing the coded instructions for functional gene products, such as proteins (Pertea and Salzberg, 2010). To what degree a gene influences a biological function will depend on whether the gene is switched on or off and how well the gene is expressed. Thus, it is natural to talk about a gene's potential to express a hereditary function. There exist different versions of genes, and errors synch as mutations, can occur. For instance will a mutation of the tumor suppressor gene TP53, highly increase your risk of getting cancer (Cheah and Looi, 2001).

A gene is expressed by first making a copy of the coding sequence of DNA, called mRNA (see Figure 2.1), which later translates into protein. One option to quantify the expression level of a gene is to count the number of mRNA in a sample. More copies of the gene leads to higher amount of mRNA and more protein is produced. The amount of proteins influences the final biological function.

An important question regarding gene expressions is, which characteristics will influence the amount of expression. This is the core of epigenomics, where the Greek term *epi* means *over, outside of* or *around*. The epigenomic has been defined in several ways (Schones and Zhao, 2008; Ptashne, 2007), but this thesis follows the definition of (Goldberg et al., 2007). Here epigenomic is the study of any potentially stable and ideally heritable change in the gene expression or cellular phenotype which is not caused by changes in the underlying DNA sequence. Epigenetic changes occur on chromatin level. Chromatin is the structure formed when DNA is wrapped around repeated histone proteins (Schones and Zhao, 2008). The functions of chro-

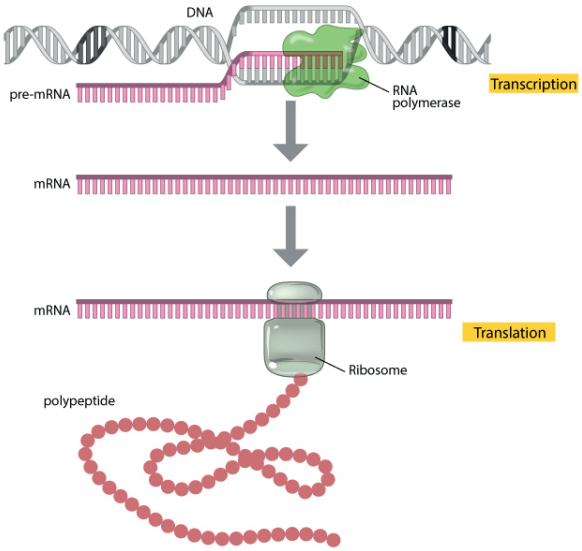

Figure 2.1: A gene is expressed through the processes of transcription and translation. During transcription, the enzyme RNA polymerase (green) uses DNA as a template to produce a pre-mRNA transcript (pink). The pre-mRNA is processed to form a mature mRNA molecule that can be translated to build the protein molecule (polypeptide) encoded by the original gene. Figure and figure text are reprinted from Nature Education (2010).

matin are many: it controls how to package DNA, avoid DNA damage, control replication, and gene expression (Francis, 2011). The spatial structure of chromatin can strongly influence the regulation of genes (Kleinjan and van Heyningen, 2005; West and Fraser, 2005). Another important epigenetic factor is DNA-methylation, where small molecules become attached to DNA. The epigenetic factors DNA-methylation and the spatial chromatin structure are said to be inheritable (Li et al., 1993; Bird, 2002; Margueron and Reinberg, 2010; Smith et al., 2002), and they are further explained in the next subsections.

2.2 The spatial structure of chromatin

The total length of the DNA strand in a human cell is about 2 meters long if it could be stretched out (Calladine et al., 2004). It therefore needs to be tightly packed to fit within the cell nucleus. It could potentially be packed to a tight ball with diameter 2 micrometers, but in practice it is less dense since it also needs to be accessible (Calladine et al., 2004). The overall chromatin structure is characterized as open or closed compartments (Lieberman-Aiden et al., 2009). The open parts (euchromatin) are associated with potentially active genes and are accessible and typically non-methylated. The closed parts (heterochromatin) often contain inactive genes, or no genes at all. These regions are highly condensed and inaccessible and usually methylated (Quina et al., 2006). Figure 2.2 shows an overview over the typical epigenetic factors, such as the spatial chromatin structure, DNA-methylation and histones. Different histones influence the structure of chromatin and also work as important epigenetic factors, but they will not be a topic

of this thesis.

For a gene to be expressed, DNA makes a loop and is able to spatially co-localize two regions called enhancers and promoters, in combination with some enzymes and proteins. The central question and a main theme of this thesis is whether or not a selected pair of enhancer and promoter regions are co-localized in 3D, or whether a group of genomic sites forms a closed region.

To map the 3D structure of chromatin advanced technology is required. The technique used in this thesis is called Hi-C (a chromosome confirmation capture technique). It uses formaldehyde to cross-link DNA, cuts it into short fragments and uses next-generation sequencing to determine the frequency of contacts between paired fragments (Lieberman-Aiden et al., 2009). The result is a contact frequency matrix over regions along DNA, aggregated over millions of cells in the same tissue. The term intrachromosomal contacts will refer to regions on the same chromosome, and interchromosomal contact refers to regions on different chromosomes. Other known techniques for 3D mapping of chromatin are 3C, 4C, 5C (Dekker et al., 2002; Simonis et al., 2006; Dostie et al., 2006) and ChIA-PET (Fullwood et al., 2009).

2.3 DNA-methylation

When methyl molecule, with the chemical structure CH_3 (one carbon and three hydrogen atoms) gets attached to a cytosine nucleotide followed by a guanine nucleotide (referred to as a CpG site), the site becomes DNA-methylated. DNA-methylation is a critical regulator of gene expression (Goldberg et al., 2007). Typically, if the promoter region of a gene is methylated, the gene is not expressed. Figure 2.3 shows that the enzyme RNA polymerase is unable to connect to the gene when it is methylated, thus the gene can not be expressed. But the association between DNA-methylation and gene expression is complex (Bird, 2002; Schones and Zhao, 2008; Wagner et al., 2014). Later sections will give a detailed look at all possible associations.

Illumina 450K bead chip is a popular platform to detect DNA-methylations and detects about 450,000 methylation sites, when're each site corresponds uniquely to a CpG location. A tissue sample will include a large number of cells, where a particular site may not be methylated on some cells, while methylated in other cells. Using the definition by Bibikova et al. (2006), the amount of DNA-methylation at a particular CpG site is reported as a proportion, where 0 means that none of the cells are methylated, and 1 means that all cells are methylated in the particular CpG. The relative score, called the beta-value, is calculated as

$$\frac{max(M,0)}{max(U,0)+max(M,0)+100}$$

where U is the fluorescent signal from the unmethylated versions and M is the signal from the methylated versions. The fluorescent signal can be negative, due to background subtraction, and is then replaced by the zero value. The constant 100 is to adjust for situations where both M and U values are small.

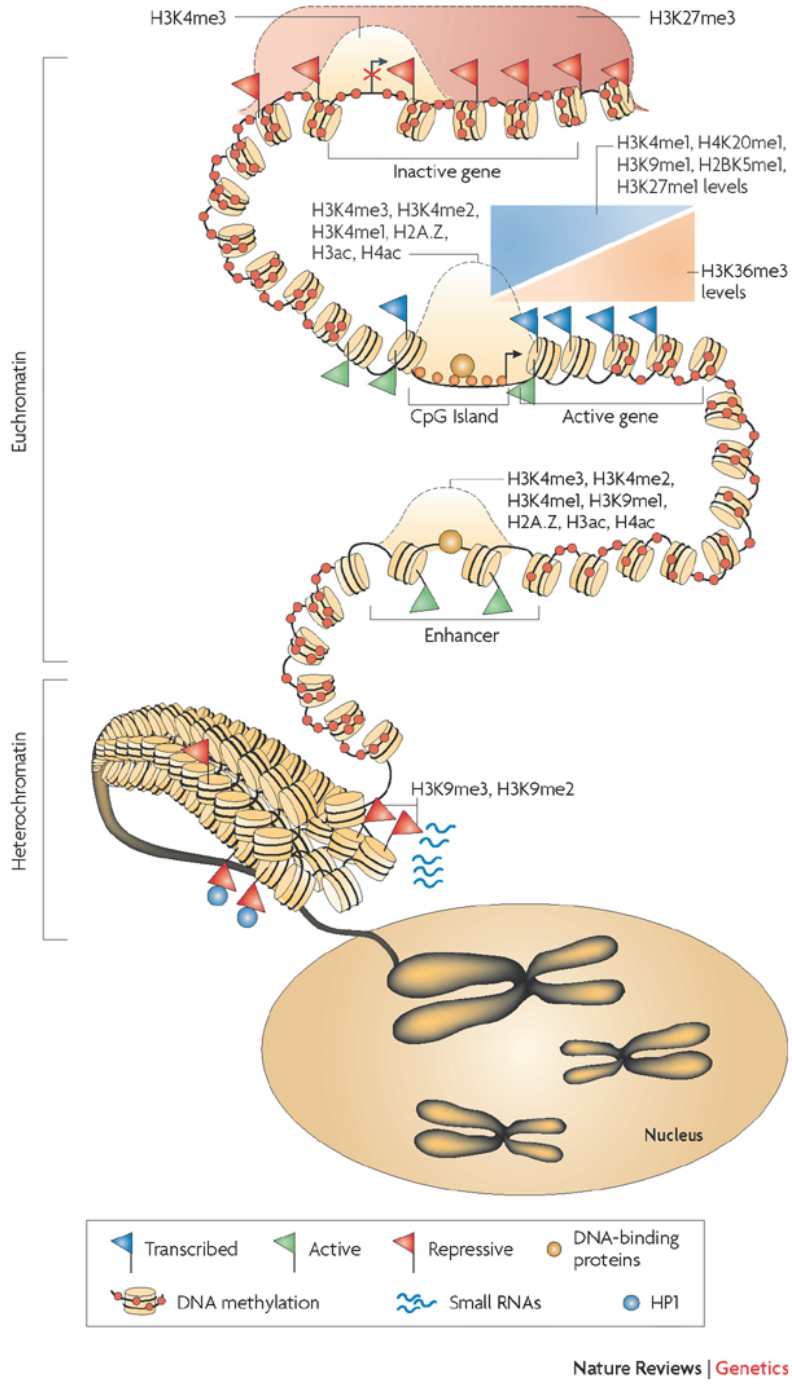

Figure 2.2: The characteristics of epigenomes includes DNA-methylation, histone modifications (for instance H3K4me3, H3K4me2) and other factors such as small RNAs, which contribute to an overall epigenome that regulates gene expression and allows cells to remember their identity. Chromatin is divided into accessible open regions called euchromatin and poorly accessible closed regions called heterochromatin. DNA-methylation is persistent throughout the genome. Figure is reprinted from Schones and Zhao (2008).

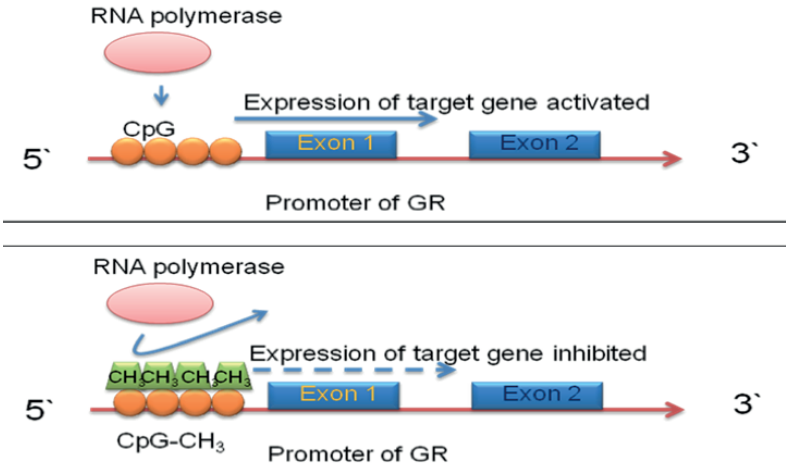

Figure 2.3: Epigenetic modification of the glucocorticoids receptor (GR) promoter by CpG altered DNA-methylation. Figure is reprinted from Erhuma (2012).

3 Aims of the thesis

By studying epigenetic markers we wish to get a better understanding of gene expressions and other biological functions. The main goal of this thesis is to develop new relevant statistical methodology for epigenomic data, which should be made publicly available as user-friendly tools. The thesis is divided into two parts, where the first part focuses specifically on hypothesis test for 3D chromatin structure and the second part on integration of genome wide data sets in penalized regression, particularly DNA-methylations together with gene expressions or other types of co-data.

In more detail, the aims of the first part are to:

- Identify characteristic properties for the joint distribution of contact frequencies in Hi-C data, evaluate the expectation, variation and correlation structures among contact frequencies, and describe the genome wide heterogeneity of the data.
- Incorporate these properties into a permutation test for 3D co-localization of genomic elements, present a proper permutation scheme reflecting the meaningful level of randomness in the data, and develop a suited effect size for the level of 3D co-localization.
- Present the proposed methodology as a publicly available tool on a web-site.

The aims of the second part are to:

- Develop integrative methodology in penalized regression, by allowing different types of co-data to be included. The penalization method should cover models both assuming and not assuming sparsity among the important covariates. The integration method should

allow for the co-data not to be included, if it does no improve the predictive performance of the method. R packages should be presented.

- Get a wider understanding of the possible associations between genome wide DNA-methylations and gene expressions.
- Apply the new integration methodologies on a data set of bone mineral density in post-menopausal women, by integrating both genome wide DNA-methylations and gene expressions.

4 Methodology

Two of the most fundamental concepts in statistics are hypothesis testing and regression. Hypothesis testing assesses evidence "in favor" or "at expense" of some claim about a phenomenon. Regression analyses evaluate specific relationships among variables. This chapter gives the background material for parametric and permutation-based hypothesis testing, where both types of tests have been proposed for the analysis of Hi-C data. For genomic and epigenomic data, multiple hypothesis tests are often performed, and we need to correct for the number of tests. It is also possible to formulate a test with a high dimensional alternative. Both these situations will be given extra attention in this chapter. If regression is the goal of the analysis, penalized regression is a popular tool and will be introduced in the second and last subsection.

4.1 Hypothesis testing

A hypothesis test is a formal procedure for studying a phenomenon with variation. With some uncertainty, one can say that the phenomenon has a specific property, either belonging to a predefined null hypothesis or an alternative hypothesis. The result from a statistical test is called significant if it is unlikely that your observations have happened by chance. There is a difference between parametric tests, where the distribution of the random variable is known, and permutation tests where no distribution for the data is assumed. The latter case is the method used in paper I in this thesis.

4.1.1 Parametric test versus permutation test

Let X be the random variable of interest with distribution $Pr(X \leq x) = F(x)$, depending on some unknown parameters θ. For simplicity, scenarios where θ is one real-valued parameter is considered. There are n independent observations, x_i for $i = 1, \ldots, n$, coming from the same distribution F. In a parametric test, a simplest decision for the parameters θ is

$$H_0 : \theta = \theta_0,$$

$$H_A : \theta \neq \theta_0,$$

for some predefined value θ_0. Let $T = T(X_1, \ldots, X_n)$ be some suitable test statistic and c the critical value. The null hypothesis is rejected if our observed test statistic $t = T(x_1, \ldots, x_n)$ exceeds c, $t \geq c$.

Two types of errors can be made. Either $t \geq c$, when in fact H_0 is true, or $t < c$, when in fact H_0 is not true. The significance level α is the probability of a type I error (Lehmann, 1999)

$$\alpha = Pr(T \geq c | H_0 \text{ true}),$$

and c is chosen such that α is equal to or lower than some predefined significance level, typically $\alpha = 0.05$.

When H_0 is true, the test statistic T will follow some distribution $Pr(T \leq t | H_0 \text{ true}) = G(t)$ depending on the distribution of X. The formal definition of a p-value is $p \equiv Pr(T > t | H_0 \text{ true})$, such that $p = 1 - Pr(T \leq t | H_0 \text{ true}) = 1 - G(t)$. It follows that the p-value is uniformly distributed over the interval $[0, 1]$, because

$$\begin{aligned} Pr(P \leq p | H_0 \text{ true}) &= Pr(1 - G(T) \leq 1 - G(t) | H_0 \text{ true}) \\ &= Pr(G(T) > G(t) | H_0 \text{ true}) = Pr(T > t | H_0 \text{ true}) \equiv p. \end{aligned} \tag{4.1}$$

This holds as long as $G(\cdot)$ is invertible and X is continuous (Murdoch et al., 2008).

The power β of a particular test is a function of the type II error and equals

$$\beta = 1 - Pr(T \leq c | H_0 \text{ not true}) = Pr(T > c | H_0 \text{ not true}). \tag{4.2}$$

If the hypothesis test is performed several times, the power will be the ratio of correctly rejected H_0 (Lehmann, 1999).

Permutation tests

In a non-parametric setting, one does not want to assume any family of distributions for the test statistic. A highly popular solution to this problem is the permutation tests, also known as resampling or randomization tests. In such type of methods, the observations are permuted to find new randomized test statistic and thus its distribution. The method is flexible and robust to missing data. Permutation tests are exact, in the sense that no assumptions or approximations are made for the distribution of the test statistic. For large samples, the parametric and permutation approaches are equivalent (Good, 2005).

The only assumption needed in a permutation test is that the resampled observations are exchangeable. The joint distribution for resample t_1, t_2 and t_3 should be independent of the ordering, such that $F(t_1, t_2, t_3) = F(t_1, t_3, t_2)$. Any order of a finite number of samples should be equally likely. Independent, identically distributed variables are always exchangeable (Good, 2005).

Let t_b be the resampled test statistic from randomizing the observations, where $b = 1, \ldots, B$.

The p-value can then be defined as in Phipson and Smyth (2010)

$$p = \frac{\left[\sum_{b=1}^{B} I(t_b > t)\right] + 1}{B+1},$$

where $I(\cdot)$ is the identity function.

There are several challenges using a permutation test. The null hypothesis needs to be precise without specifying a distribution with corresponding parameters for the data. The randomization scheme must reflect the desired null distribution for the test statistic, such that the test actually has significance level α. This means that if the test is performed several times when H_0 is true, the null hypothesis should be rejected $\alpha \times 100\%$ of the cases. One way to check this is to simulate data where H_0 is in fact true and run the hypothesis test. If the test is correctly specified, the p-values should be uniformly distributed when H_0 is true, as seen in Equation (4.1).

Another challenge is to find the power of the test. This can be difficult due ti the unknown distribution of the test statistic under the alternative hypotheiss and will be discussed in the Discussion.

4.1.2 Hypothesis testing on 3D chromatin structure

When analyzing Hi-C data, one is typically interested in a predefined set of sites along the genome, and the co-localization between those elements. One of the first methods to handle this question was proposed by Botta et al. (2010). A permutation test was suggested, where the 3D contacts were considered as independent, and uniformly randomized. Duan et al. (2010) and Dai and Dai (2010) presented a parametric test by assuming that the 3D contacts were hypergeometrically distributed, thus again assuming that the 3D contacts are independent. When evaluating the 3D structure, it is clear that some of the contacts are extremely correlated. For instance, the transitivity relation gives that if a element i and j are close to k, then also i and j are close. Witten and Noble (2012) analyzed interchromosomal 3D contacts and accounted for the dependencies in the 3D structure, by proposing an uniform resampling of the position of the genomic elements, and thus letting the 3D data itself be fixed. Paper I will go far beyond this paper in treating different forms of dependencies in the data. In the presented literature, the 3D contacts are often referred to as 3D interactions. Since this term may be confusing in statistics, use the term 3D contacts is instead used.

Additional properties need to be considered when analyzing intrachromosomal 3D contacts. The genomic distance between the elements is the linear distance along the genome, often counted as number of base pairs along the DNA strand. For intrachromosomal 3D contacts, a higher contact frequency is expected between elements with short genomic distance.

Also, the distribution of the 3D contact frequencies varies throughout the genome, and is not homogeneous. For instance, it depends on whether the genomic elements are close to the center (centromeres) or the end (telomeres) of the chromosome, and the GC content (the intensity of C and G nucleotides) in the region. It should be possible to take into these properties in a hypothesis test.

4.1.3 Controlling the false discovery rate in multiple testing

When multiple tests are being performed, the overall false discovery rate (the rejection of H_0 when it is actually true) needs to be controlled. Benjamini and Hochberg (1995) proposed a method for adjusting the p-values with respect to the overall false discovery rate (FDR) among m hypotheses. There are $i = 1, \ldots, m$ hypotheses on the form

$$H_{0i} : \theta_i = \theta_{0i} \quad H_{Ai} : \theta_i \neq \theta_{0i}.$$

The un-adjusted p-values, p_i, are calculated by some method for each i individually. The number of rejected and accepted hypotheses are shown in Figure 4.1, where V, U , T, S, and m_0 are unknown quantities. V is the number of type I errors, and T is the number of type II errors. One is prepared to tolerate some type I errors, provided their number is small in comparison to R, the total number of significant tests. FDR is defined as the expected proportion of type I errors among the rejected hypotheses, $E(V/R)$, which is put equal to zero when $R = 0$. The procedure to control this quantity was presented by Benjamini and Hochberg (1995). It holds under certain dependency structures including positive dependences between the tests (Benjamini and Yekutieli, 2001) .

Number of errors committed when testing m null hypotheses

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	$m - m_0$
	$m - \mathbf{R}$	**R**	m

Figure 4.1: Figure reprinted from Benjamini and Hochberg (1995).

4.1.4 A test for a high dimensional alternative

A utility test developed by Goeman et al. (2006) can be used to test against a high dimensional alternative. This test, called the global test, allows for the alternative hypothesis to be a high-dimensional model. Suppose that the outcome variable Y has distribution dependent on p variables $x_1, \ldots, x_p$ through some coefficients $\beta = (\beta_1, \ldots, \beta_p)$. The test is formulated as $H_0 : \beta = 0$ against $H_A : \beta \neq 0$. The null hypothesis is that there is no association between Y and the p variables, with alternative hypothesis that one or more variable is associated with Y. A prior distribution is assumed for β, where $\mathrm{Cov}(\beta) = \tau^2\Sigma$, and the test can be written as $H_0 : \tau^2 = 0$ against $H_A : \tau^2 > 0$. For the linear model, Goeman et al. (2006) showed that the test has good power, even when compared to the classical tests in low dimensions.

4.2 Regression in high dimensions

This section describes regression modeling, by first introducing ordinary regression where we have less variables than samples, so called $p \leq n$. In the opposite scenario, $p > n$, the penalized methods lasso and ridge regression will be presented. Lastly, we introduce penalized methods with multiple penalty terms.

4.2.1 Ordinary regression $(p \leq n)$

Let there be n independent observations of some outcome y_i and variables $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ for $i = 1, \ldots, n$. A regression model analyzes the relationship between the outcome and the p variables or predictors. The following theory on generalized linear models is based on the book by Dobson and Barnett (2008).

Generalized linear models (GLM)

Based on some unknown coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, we assume a true linear relationship between the predictors and some monotone link function $g()$ of the expected value of the random outcome Y_i. The model is

$$g(E(Y_i)) = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

For reasons of simplicity, the data is mean centered, such that the intercept $\beta_0 = 0$, and the number of coefficients is p. On matrix form the model is written $g(E(\boldsymbol{Y})) = \boldsymbol{X}\boldsymbol{\beta}$ for $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{X}$ the $n \times p$ design matrix with rows $\boldsymbol{x}_i^T$.

In this type of generalized linear model, the distribution of Y belongs to the exponential family, for instance binomial (logistic regression), Poisson or Gaussian. McCullagh and Nelder (1989) showed how to convert the proportional hazard model in survival analysis to also fit the generalized linear models.

Maximum likelihood estimate (MLE)

Estimates of $\boldsymbol{\beta}$ are found by maximizing the log likelihood, $l(\boldsymbol{\beta})$. It is $\boldsymbol{\beta}$ where the derivative of the log likelihood is zero. For many of the distributions at hand, there are no explicit formula for $\boldsymbol{\beta}$ in this case. A typical methods to use is then the Newton-Raphson iteration,

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \left[l''(\boldsymbol{\beta}^{(k)}) \right]^{-1} l'(\boldsymbol{\beta}^{(k)}).$$

The procedure is iterated until the difference between $\boldsymbol{\beta}^{(k+1)}$ and $\boldsymbol{\beta}^{(k)}$ is smaller than some threshold. The solution of $\boldsymbol{\beta}$ can be written as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_{\boldsymbol{w}}^{\boldsymbol{T}} \boldsymbol{X}_{\boldsymbol{w}})^{-1} \boldsymbol{X}_{\boldsymbol{w}}^{\boldsymbol{T}} \boldsymbol{z}, \tag{4.3}$$

for appropriate $\boldsymbol{X}_{\boldsymbol{w}} = \boldsymbol{W}\boldsymbol{X}$, where W is some weights, and $\boldsymbol{z} = (z_1, \ldots, z_n)$ corresponding to the distribution at hand (Dobson and Barnett, 2008, Chapter 4, page 63-64). For instance, in logistic regression with the logic link function we have that $\boldsymbol{W} = \text{diag}(\sqrt{p_1(1-p_1)}, \ldots, , \sqrt{p_n(1-p_n)})$

for $p_i = \text{expit}(\boldsymbol{x}_i^T\boldsymbol{\beta})$, and $z_i = \text{logit}(p_i) + \frac{y_i - p_i}{p_i(1-p_i)}$. Note that both $\boldsymbol{X_w}$ and $\boldsymbol{z}$ depends on $\boldsymbol{\beta}$.

In linear regression, the matrix $\boldsymbol{X_w}$ is simply the design matrix $\boldsymbol{X}$ and $\boldsymbol{z} = \boldsymbol{y}$, such that the estimate of $\boldsymbol{\beta}$ becomes $\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$.

Validation of prediction models

After estimating the coefficients of a model, the predictive performance for new data is important to evaluate. If the model only fits the current data, it is not a desirable model. To evaluate the predictive performance one needs training data $(\boldsymbol{y}_{\text{train}}, \boldsymbol{X}_{\text{train}})$ to train the model $\hat{\boldsymbol{\beta}}_{\text{train}}$, and test data $(\boldsymbol{y}_{\text{test}}, \boldsymbol{X}_{\text{test}})$ to predict the outcome $\hat{\boldsymbol{y}}(\boldsymbol{X}_{\text{test}}, \hat{\boldsymbol{\beta}}_{\text{train}})$. For the linear case the predicted outcome becomes $\hat{\boldsymbol{y}} = \boldsymbol{X}_{\text{test}} \times \hat{\boldsymbol{\beta}}_{\text{train}}$. Finally, $\hat{\boldsymbol{y}}$ is compared to the truth $\boldsymbol{y}_{\text{test}}$. For logistic regression we model the predicted probability of the event, and for survival we model the predicted event probability at different times.

The test data should be some new independent comparable data set. If this is not possible to obtain, cross-validation of the one data set at hand is a common approach. In a cross-validation setting, the data is divided into K partitions, where $1/K$ is used as test data and the other proportion is used to fit the model. This procedure is done K times for all possible combinations of training and test data.

It is a variety of different methods used to measure the predictive performance of a fitted model. Some of the most common ones are briefly described next.

The predictive mean square error $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is often reported. For logistic regression, this is called the Brier score and measures the difference between the true outcome and the predicted probability of the event (Brier, 1950; Pepe et al., 2004). For survival, the Brier score includes some weights, taken into consideration the conditional probability of being uncensored during time (Graf et al., 1999).

A receiver operating characteristic (ROC) curve shows the model's ability to discriminate between two groups based on different thresholds. The curve shows the sensitivity and specificity of the binary classifier. Often the area under the ROC curve (AUC) is reported. When AUC is equal to 0.5, the decision made by the model is no better than a random guess. For survival, the time-dependent ROC curve is used (Heagerty et al., 2000).

4.2.2 Penalized regression $(p > n)$

When there are fewer variables than observations ($p \leq n$), Equation (4.3) has an unique solution. In contrast, when $p > n$, $\boldsymbol{X_w^T X_w}$ is singular and noninvertible, there are infinitely many solutions and penalized regression is needed.

Penalized likelihood

In penalized regression, a restriction is enforced on the size of the coefficients. To this end, one introduces a penalty function $J(\boldsymbol{\beta})$ and maximizes the log likelihood function under the restriction that $J(\boldsymbol{\beta}) < s$ for some chosen value s. The consequence of this is that the estimated coefficients are shrunken towards zero. The maximum of the log likelihood with respect to the

restriction can be written, on the Lagrangian form, as

$$\hat{\boldsymbol{\beta}}^{\lambda} = \max_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta}) \right\}, \tag{4.4}$$

for some penalty parameter $\lambda \geq 0$. The solution space for larger value of $s \in (0, \infty)$, is equal to the solution space of the Lagrangian form for smaller values of λ. Larger values of λ lead to more shrinkage towards zero in the $\hat{\boldsymbol{\beta}}$. This leads to more bias and less variance (Hastie et al., 2009; Bühlmann and van de Geer, 2011). The optimal value of λ can therefore be chosen such that the bias-variance tradeoff is optimal, for instance by minimizing the predictive mean square error (pMSE) by use of K-fold cross validation. Typically $K = 10$, so that nine tenth of the data serve as training set, and $1/10$th as test data.

However, Equation (4.4) depends on the scaling of the data. Therefore, normalization of the data is typically done by centring each covariate so they have mean zero and standard deviation one.

The lasso

The least absolute shrinkage and selection operator (lasso) was first introduced by Tibshirani (1996). The penalty term is based on the $L1$-norm $J(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$, favoring sparse solutions where a large proportion of the coefficients are exactly zero. The lasso can select at most n variables out of p candidates (Efron et al., 2004).

There are versions of the lasso that posses the oracle property. This means that the methods find estimators that are asymptotically as good as the oracle estimator, which a priori knows the non-zero β. There are three assumptions that need to be fulfilled. First, there should only be weak correlation between predictors, especially between relevant and irrelevant covariates. Secondly, the true model must be highly sparse with less influential covariates than n. Thirdly, the minimum value of β must be such that $min_i(|\beta_i| : \beta_i \neq 0) \geq K$, for some constant $K > 0$ (Bühlmann and van de Geer, 2011).

In the case of orthonormal $\boldsymbol{X}$ in linear regression, there exists a closed form solution for the lasso, namely the soft thresholding $\hat{\beta}_j^{\lambda} = sign(\hat{\beta}_j)(\hat{\beta}_j - \lambda)^+$, for the least square solution $\hat{\beta}_j$ (Hastie et al., 2009).

Ridge regression

Ridge regression was introduced for the linear model by Hoerl and Kennard (1970), and it is based on the L2-norm with penalty $J(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_2^2 = \sum_{j=1}^{p} \beta_j^2$. It favors solutions with many small non-zero coefficients, and does not do variable selection.

In Figure 4.2, we see a illustration of lasso and ridge regression for $p = 2$. Here, β_1 is on one axis and β_2 on the other. The ellipses represent the log likelihood with maximum value in the center, the black square shows the lasso constraint ($|\beta_1| + |\beta_2| < s$) and the black circle the ridge regression constraint ($\beta_1^2 + \beta_2^2 < s$). We see that in the lasso case the maximum of the log likelihood together with the L1-norm restriction often lead to one of the variables being exactly zero, which is not the case for the ridge regression. This is why the lasso results in a selected list of coefficients different from zero, in contrast to ridge regression, which will include all p

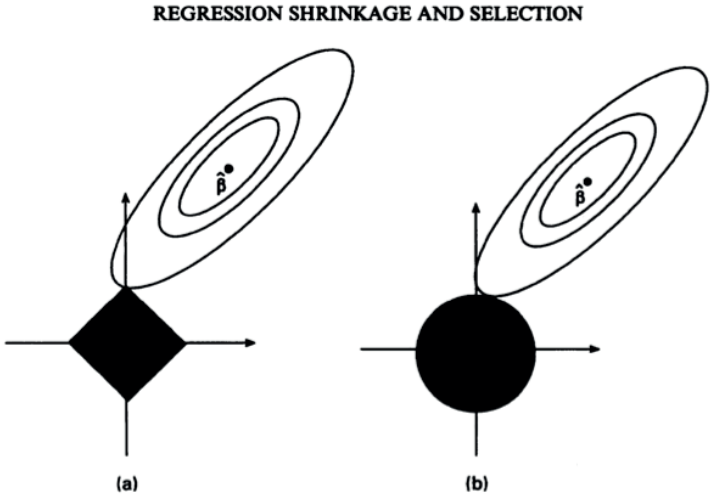

Figure 4.2: Estimation picture of (a) the lasso and (b) ridge regression. Figure reprinted from Tibshirani (1996).

coefficients.

More details on the ridge regression solution

When using the ridge penalty we can find a closed-form solution for $\hat{\beta}$. With a closer look at the ridge regression solution we see some interesting properties for the final estimates. The following calculations are based on a paper by Le Cessie and Van Houwelingen (1992).

We start with the penalized log likelihood for the ridge regression $l^{\lambda}(\beta) = l(\beta) - \lambda\beta^T\beta$. Let the first and negative second derivative of the log likelihood $l(\beta)$ be noted as U and Ω, respectively. For the penalized log likelihood we then obtain $U^{\lambda}(\beta) = U(\beta) - 2\lambda\beta$ and $\Omega^{\lambda}(\beta) = \Omega(\beta) + 2\lambda$. Similarly as for the non-restricted case we can find the Taylor series expansion of the first derivative U^{λ} around the true value β_0 as

$$U^{\lambda}(\beta) = U^{\lambda}(\beta_0) - \Omega^{\lambda}(\beta_0)(\beta - \beta_0) + o(||\beta - \beta_0||).$$

Using the fact that $U^{\lambda}(\hat{\beta}^{\lambda}) = 0$ we see that

$$\hat{\beta}^{\lambda} \approx \beta_0 + [\Omega(\beta_0) + 2\lambda]^{-1}(U(\beta_0) - 2\lambda\beta_0) = [\Omega(\beta_0) + 2\lambda]^{-1}(\beta_0\Omega(\beta_0) + U(\beta_0)).$$

The unrestricted estimates can be approximated as $\hat{\beta} \approx \beta_0 + \Omega(\beta_0)^{-1}U(\beta_0)$, so the estimate of $\hat{\beta}_{\lambda}$ can be written

$$\hat{\beta}^{\lambda} \approx [\Omega(\beta_0) + 2\lambda I]^{-1}\Omega(\beta_0)\hat{\beta}, \tag{4.5}$$

where $\Omega(\beta_0)$ is typically approximated by $\Omega(\hat{\beta})$. For linear regression, this expression simpli-

fies to

$$\hat{\boldsymbol{\beta}}_\lambda = [\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I}]^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

For orthonormal $\boldsymbol{X}$, the expression further simplifies to $\hat{\boldsymbol{\beta}}_\lambda = \frac{1}{1+2\lambda}\hat{\boldsymbol{\beta}}$, and we see that the ridge estimate is a simple shrinkage of the ordinary least squares estimate.

Simplify calculations by the use of singular value decomposition

The singular value decomposition (SVD) is extremely useful for calculating the ridge solution. To this end we write the $n \times p$ matrix $\boldsymbol{X}$ as $\boldsymbol{UDV}^T$, where $\boldsymbol{U}$ $(n \times p)$ and $\boldsymbol{V}$ $(p \times p)$ are orthonormal matrices spanning the column and row space of $\boldsymbol{X}$. The $p \times p$ matrix $\boldsymbol{D} = \text{diag}(d_i)$ is a diagonal matrix with the singular values d_i of $\boldsymbol{X}$ on the diagonal. Remember that $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$. Then

$$\begin{aligned}
\boldsymbol{X}^T\boldsymbol{X} &= (\boldsymbol{UDV}^T)^T\boldsymbol{UDV}^T = \boldsymbol{VDU}^T\boldsymbol{UDV}^T = \boldsymbol{VD}^2\boldsymbol{V}^T \\
[\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}]^{-1} &= [\boldsymbol{VD}^2\boldsymbol{V}^T + \lambda\boldsymbol{I}]^{-1} = [\boldsymbol{V}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})\boldsymbol{V}^T]^{-1} = \boldsymbol{V}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T \\
&= \boldsymbol{V}\text{diag}\left(\frac{1}{d_i^2 + \lambda}\right)\boldsymbol{V}^T \\
[\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}]^{-1}\boldsymbol{X}^T &= \boldsymbol{V}\text{diag}\left(\frac{1}{d_i^2 + \lambda}\right)\boldsymbol{V}^T\boldsymbol{VDU}^T = \boldsymbol{V}\text{diag}\left(\frac{d_i}{d_i^2 + \lambda}\right)\boldsymbol{U}^T.
\end{aligned}$$

Similar calculations are valid for generalized linear models using $\boldsymbol{X}_{\boldsymbol{w}}$ instead of $\boldsymbol{X}$.

R packages

There are several R packages that may be used for lasso and ridge regression. Two of them are *glmnet* by Friedman et al. (2010) and *penalized* by Goeman (2010). There are several differences between these two packages as seen in Table 4.1. In practice, one should keep in mind when comparing the optimal λ, that different scaling is used for the objective function for the two packages, which leads to a transformation of λ. Further, the optimization procedure is different, together with different loss functions in the cross validation, but this only results in small differences in the optimal value of λ. For the overall comparison the running time is often the critical part. Both methods are fast, but often glmnet is fastest, which is shown for Cox regression in Simon et al. (2011).

4.2.3 Penalized regression with multiple penalties

In the penalized methods presented so far, all regression coefficients are penalized with one common parameter so all covariates are treated equally. If multiple penalty terms, $\lambda_j = \lambda l_j$ is introduced, the penalized log likelihood can be written as

$$l^\lambda(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \sum_{i=1}^{p} \lambda_j|\boldsymbol{\beta}_j|^r = l(\boldsymbol{\beta}) - \lambda\sum_{i=1}^{p} l_j|\boldsymbol{\beta}_j|^r,$$

where $r = 1$ is the lasso, and $r = 2$ is ridge regression. The penalty terms can be different for all covariates or groups of covariates. A wide range of methods emerges, particularly based on the lasso, and a selection of these methods are presented next.

Characteristics	**glmnet**	**penalized**
Objective function	$n^{-1}l(\beta) - \lambda J(\beta)$	$l(\beta) - \lambda J(\beta)$
The λ path	Preselected grid of length 100, $\hat{\beta}(\lambda_{max}) = 0$, $\lambda_{min} = 0.001 \times \lambda_{max}$	Iterates to the optimal loss function
Optimizing scheme	Coordinate decent (Friedman et al., 2010), solving the problem along the preselected path of values for λ, using the current estimates as warm starts.	Follows the gradient of the log likelihood from a given starting value of β. Automatically switch to a Newton-Raphson algorithm when it gets close to the optimum
Loss function	Mean squared (absolute) error, deviance, partial-likelihood, misclassification error, area under the ROC curve.	Cross-validated likelihood
K-fold CV	Default is 10	Default is n

Table 4.1: The differences between the two R-packages glmnet and penalized used to maximize the penalized log likelihood.

Adaptive lasso

In adaptive lasso, presented by Zou (2006), the penalty weights are defined as $l_j = 1/\hat{\beta}_j^L$, where $\hat{\beta}_j^L$ is the estimate from the classical lasso. The covariates that classical lasso finds important, are given a lower penalty in the adaptive lasso, while those set to 0 in the first iteration are no longer included. The result is that fewer variables are selected, and their coefficients are hardly shrunken. Under some assumptions on the design matrix $\boldsymbol{X}$ and sufficiently large non-zero coefficients, the adaptive lasso possess oracle properties, meaning that the method can correctly select the nonzero coefficients with a probability converging to one.

General multi-penalty

The covariates are divided into predefined groups, and one penalty term λ_g are introduced per group $g = 1, \ldots, G$ (Tai and Pan, 2007). If the number of groups is small, each λ_g is found by cross validation, which means in total a G-dimensional cross validation. For large G this may be computationally too demanding, so it is suggested to use $\lambda_g = \lambda l_g$, and only determined λ by cross validation. The penalty weights l_g is then the mean of the absolute value of the variables in group g.

Group lasso

For group lasso, predefined groups are included, such that either all covariates in a group are selected or none of them are selected. The penalty is a L1-norm at the group level, and a L2-norm for the features within each group, $\lambda \sum_{l=1}^{L} \sqrt{p_l} ||\beta^{(l)}||_2$, where $\beta^{(l)}$ is the coefficient vector of group l, and p_l is the length of $\beta^{(l)}$ (Yuan and Lin, 2006). Later, the sparse group lasso was presented by Simon et al. (2013) as $(1-\alpha)\lambda \sum_{l=1}^{L} \sqrt{p_l} ||\beta^{(l)}||_2 + \alpha\lambda ||\beta||_1$, where $\alpha = 0$ gives the group-lasso and $\alpha = 1$ gives the classical lasso. This method results in sparsity within the selected groups.

Weighted lasso with data integration

To incorporate external knowledge of believed relevance of each covariate, Bergersen et al. (2011) presented the weighted lasso with data integration. Each individual penalty weight is calculated based on some co-data together with an additional introduced parameter to decide upon the relative strength of the external weights. The method is data driven, so the data decides the range of the individual penalties through a two-dimensional cross-validation. If the co-data is not relevant, the method allows for no weights.

Garcia et al. (2013) further build upon this method, where they, based on some external information, divide the covariates into two groups. The first group of covariates should not be subject to selection, and the second group of covariates gets individual penalties such that some of them are more likely to be in the model than others. The penalties for the first group are constructed such that those covariates are guaranteed to be in the model.

4.2.4 Bayesian perspective

Lasso and ridge regression have a Bayesian interpretation. If a prior for $\boldsymbol{\beta}$ is defined as a Gaussian $N(0,\tau^2)$, the mode (and the mean) of the corresponding posterior distribution is the ridge estimate, where $\lambda \propto 1/\tau^2$. If the prior for $\boldsymbol{\beta}$ is defined as double-exponential $(\tau/2)^p \exp(-\tau||\boldsymbol{\beta}||_1)$, the mode of the posterior is the lasso estimate (Hans, 2009). The idea of individual penalties could therefore be viewed as having different priors in terms of the variances τ_j^2 for each coefficient $\boldsymbol{\beta}_j$.

Empirical Bayes approach is when the hyperparameters (for instance τ or τ_j) are estimated from the data itself, and plugged into the posterior (Efron and Morris, 1973; Morris, 1983). Typically, an estimate of the hyperparameter is found by maximizing the marginal distribution over the hyperparameter. A simple example is for a Gaussian model $y_i|\theta_i \sim N(\theta_i,\sigma^2)$ (σ^2 known) and a Gaussian prior $\theta_i \sim N(\mu,\tau^2)$ for $i = 1,\ldots,n$. Then the marginal $m(y_i|\mu,\tau^2) \sim N(\mu,\sigma^2+\tau^2)$, and when maximizing it as a function of (μ,τ^2) we get $\hat{\mu} = \bar{y} = \sum_i y_i$ and $\tau^2 = (s^2-\sigma^2)^+ = \max(0, s^2-\sigma^2)$, where $s^2 = \frac{1}{n}\sum_i (y_i-\bar{y})^2$.

5 Summary of papers

5.1 Paper I

Jonas Paulsen*, Tonje G. Lien*, Geir Kjetil Sandve, Lars Holden, Ørnulf Borgan, Ingrid K. Glad and Eivind Hovig (2013). Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic Acids Research*, 41(10): pp 5164–5174. *Joint first authors.

Paper I introduces a novel permutation test and an effect size for spatially co-localized genomic elements. The main question addressed is whether a set of regions in the genome (our "query set of interest") is spatially closer to each other in 3D than expected by chance. The null hypothesis

assumes every contact to be random. The test includes both testing for genomic elements on the same chromosome (intra) and on different chromosomes (inter). Two genomic elements close to each other along the genome will, just by chance, often be 3D co-localized. To account for this we standardize each 3D contact frequency by their expectation and standard deviation. The overall test statistic is then the mean over all standardized 3D contact frequencies in the query set of interest.

Because of the complex joint distribution of 3D contact frequencies, we present a permutation test, permuting the position of the genomic elements of interest. The strong correlation structures within the data are taken into account by not permuting the positions uniformly, but by including certain novel restrictions. Additional properties, for instance the GC content, are also considered and can be preserved by only permuting elements with similar properties. The global 3D data itself is not randomized and is thus fully maintained.

The 3D structure under the null hypothesis is simulated by a random walk inside a reflecting sphere. The simulation studies show that the permutation test truly has significance level α. Finally, the hypothesis test is tested on publicly available biological data using different cell lines; IMR90, human embryonic stem cell and K562 (Dixon et al., 2012; Lieberman-Aiden et al., 2009). The hypothesis test and all biological data are nicely implemented in a user-friendly web page, available at https://hyperbrowser.uio.no/3d/.

5.2 Paper II

Sjur Reppe, Tonje G. Lien, Vigdis T. Gautvik, Ole K. Olstad, Hege G. Bakke, Robert Lyle, Marianne Kringen, Ingrid K. Glad and Kaare M. Gautvik (2015). DNA methylations in bone correlate markedly to BMD associated transcript levels and distinguish between osteoporotic and healthy postmenopausal women *(Manuscript)*.

Paper II investigates a selection of DNA-methylation sites and gene expressions in relation to bone mineral density (BMD) in a study of 80 women. The 100 gene expressions most associated with BMD, found by Reppe et al. (2010), and their 2529 cis related DNA-methylations are the focus. Osteoporosis (OP) is a disease associated with loss of bone mineral density, and the women in the cohort study represent a diversity of BMD including both OP cases and healthy individuals. Five DNA-methylation sites are found to be significantly differentiated (at 5% FDR) between the healthy and sick patients, when accounting for age.

To get a deeper understanding of the association between gene expression and DNA-methylation, we test for significant correlation in both cis and trans related pairs. A cis relation refers to DNA-methylation and transcript from the same gene, and trans refers to DNA-methylations and transcripts from different genes. By the use of a permutation test, we see that the correlation between DNA-methylation and gene expression can be both significantly positive and negative, and occur between both cis and trans related pairs. At a 5% FDR, 1470 pairs are significantly correlated (corresponding to, mainly, correlations larger than 0.4 or lower than -0.4). Only 5 of these 1470 significant correlations are in cis pairs. This gained knowledge is incorporated in an integrative analysis of BMD in paper IV, using the full DNA-methylation data and genome wide gene expressions as co-data.

5.3 Paper III

Mark A. van de Wiel, Tonje G. Lien, Wina Verlaat, Wessel N. van Wieringen and Saskia M. Wilting (2015). Better prediction by use of co-data: Adaptive group-regularized ridge regression. *Statistics in Medicine (Published)*.

Paper III develops one of the methods applied in paper IV, and the presented method is a ridge regression that allows for multiple group-penalties. The motivation is a global regression analysis (typically with more covariates than samples) with integrated co-data, where the groups are defined by the co-data. Covariates in the same group are believed to have similar predictive impact. The group-specific penalties are then found by the use of empirical Bayes, which adapt to the amount of information in the co-data. Therefore, if the groups do not give improved predictive performance, the groups will not be integrated in the final model. When there are ordered groups, such that some groups are favored over others, a constraint is enforced on the group-specific penalties.

Simulations and two biological examples validate the method. For both biological examples, the new method clearly improved predictive performance compared to ordinary ridge regression and group lasso. The results show that the method may lead to better distinction between "near-zero" and relatively large regression parameters, which promote post-hoc variable selection. Lastly, the method is implemented in a R package, termed GRridge, available at http://www.few.vu.nl/ mavdwiel/grridge.html.

5.4 Paper IV

Tonje G. Lien, Sjur Reppe, Kaare M. Gautvik, Ørnulf Borgan and Ingrid K. Glad (2015). Integration of epigenomic and genomic data in high-dimensional penalized regression. A cohort study on bone mineral density *(Manuscript)*.

In paper IV, the method developed in paper III is applied to the bone mineral density (BMD) data, where BMD is used as response, genome wide DNA-methylations as covariates and gene expressions as co-data. The DNA-methylations with a strong association to genome wide gene expressions are favored in the penalized regression. The p-values from the global test (described in Section 4.1.4), adjusted for multiple testing, are used to quantify the strength of association between DNA-methylations and gene expressions and form the basis for the grouping of the covariates in the group-regularized ridge. The results show a considerable improvement in prediction error with 14% compared to classical ridge regression. A relative small group of covariates was assigned small penalties, and the remaining covariates were given relatively large penalties, in practice excluding them from the final model.

Variable selection was also desirable in the analysis of DNA-methylations in the BMD study. Therefore, an improved extension of the weighted lasso with data integration is presented. Here, each individual penalty is defined to be proportional to a weight, where the weight reflects the relevance of the covariate according to the co-data. Again, the multiple corrected p-values from the global test were used as basis for the penalty weights. An additional parameter was introduced to control the relative difference between the minimum and maximum penalty weight.

The optimal value of this parameter is decided by the data itself, such that the penalty weights will all be equal to 1 if there are no information in the multiple corrected p-values. This was not the case for the BMD data, and in fact we saw a great improvement of 17% in terms of predictive performance in the weighted lasso with data integration compared to the classical lasso with no penalty weights. The R package glmnet was used to run the method, since this package has penalty weights as a separate argument in the function call.

Lastly, p-values from testing each DNA-methylation against only their cis related gene expressions were also tried as basis for the penalty weights. But this showed no improvements in prediction, for either of the two methods. These results indicate that DNA-methylations are involved with and associated to other genes than their cis related genes. Similar results was seen in paper II.

6 Methodological extensions and future work

This section presents extensions to paper I and paper III. The extensions to paper I have been published and implemented online. The extensions to paper III have been implemented in the corresponding R package and will be published later. Lastly, ongoing work is presented, where Hi-C data is being analyzed in a penalized regression setting.

6.1 Extensions to paper I

Several additional hypothesis have been developed, which were not included in paper I, but published by Paulsen et al. (2014). Alternatively, one can ask questions such as: Is one query set more/less co-localized with another query set in 3D? Are selected pairs of elements more co-localization in 3D? There are also different options for restrictions in the randomizations concerning predefined categories. For all the hypothesis tests, a representative enrichment score is given, showing the 3D co-localization compared to the expected 3D co-localization. Visualization of Hi-C data as heat maps and graphs were made available online.

6.2 Extensions to paper III

The adaptive group-regularized ridge regression handles logistic and linear regression models. But, it is fairly straightforward to extend the empirical Bayes estimate in the method to also handle Cox regression for censored survival data. For survival data, individual i has follow-up time t_i and event indicator d_i, where $i = 1,\ldots,n$. The event indicator is equal to 1 if the event of interest has occurred and equal to 0 if the individual is censored. For each individual, p explanatory variables $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ip})^T$ are measured. The design matrix is denoted $\boldsymbol{X}$, a $n \times p$ matrix with rows $\boldsymbol{x}_i^T$. Cox regression is defined by assuming that the hazard function of

the ith individual takes the form

$$h(t|\boldsymbol{x}_i) = h_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is the p-dimentional vector of coefficients.

Next, we follow the approach of van Houwelingen et al. (2006), and use the full log likelihood for the Cox model, given by

$$l(\boldsymbol{\beta}, h_0) = \sum_{i=1}^{n}\left[d_i\left\{\log h_0(t_i) + \boldsymbol{x}_i^T\boldsymbol{\beta}\right\} - H_0(t_i)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\right],$$

where the cumulative baseline hazard $H_0(t) = \sum_{s\leq t} h_0(s)$ is assumed to be a step function with jumps at the observed event times t_i. For $p > n$, the penalized full log likelihood is

$$l_{pen}(\boldsymbol{\beta}, h_0) = l(\boldsymbol{\beta}, h_0) - \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}.$$

To find the expectation and covariance of $\hat{\boldsymbol{\beta}}$, needed in the adaptive group-regularized ridge regression, we find the first and second derivative of the penalized full likelihood. The first derivative is

$$\frac{\partial l_{pen}(\boldsymbol{\beta}, h_0)}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}\left[d_i - H_0(t_i)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\right]\boldsymbol{x}_i - 2\lambda\boldsymbol{\beta} = \boldsymbol{X}^T\Delta - 2\lambda\boldsymbol{\beta},$$

where $\Delta = (\Delta_1, \ldots, \Delta_n)^T$ $\Delta_i = d_i - H_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$, and the second derivative is

$$\frac{\partial^2 l_{pen}(\boldsymbol{\beta}, h_0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -\left[\sum_{i=1}^{n} H_0(t_i)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i\boldsymbol{x}_i^T + 2\lambda\boldsymbol{I}\right] = -(\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X} + 2\lambda\boldsymbol{I}),$$

where $\boldsymbol{D} = \text{diag}\left\{H_0(t_i)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\right\}$. Then, by an argument similar to the one in Le Cessie and Van Houwelingen (1992) and Cule et al. (2011), we find the first-order approximation of $E(\hat{\beta}_k)$ as

$$\begin{aligned}\mu_k &= [\boldsymbol{\beta} - 2\lambda(\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{\beta}]_k = [\{\boldsymbol{I} - 2\lambda(\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\}\boldsymbol{\beta}]_k \\ &= [(\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\{\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X} + 2\lambda\boldsymbol{I} - 2\lambda\boldsymbol{I}\}\boldsymbol{\beta}]_k = [(\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X}\boldsymbol{\beta}]_k,\end{aligned} \tag{6.1}$$

where $[M]_k$ denotes the kth component of a vector M. Next, the estimated covariance matrix $\Sigma = \text{Cov}(\hat{\boldsymbol{\beta}})$, is approximated by

$$\Sigma \approx (\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X} + 2\lambda I)^{-1}\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X} + 2\lambda I)^{-1}. \tag{6.2}$$

In the group-regularized ridge regression, the first-order approximation of the first and second moments, Equations (6.1) and (6.2), are used in the calculations of the empirical Bayes estimator.

6.3 Future work using Hi-C data in penalized regression

A possible direction is to analyze the Hi-C data in a regression setting. For instance, the data presented in Jin et al. (2013) includes Hi-C experiments in primary human fibroblast cells (IMR90), for 6 biological replicates in two different conditions, before and after treating the samples with TNF-α signaling. For this type of data, a penalized logistic regression using the 3D contacts as covariates could be suitable. The goal is to find a list of contacts, which differs between the binary outcomes.

Hi-C data contains strong correlations and this structure should be taken info account. The fused lasso, proposed by Tibshirani et al. (2005) has the restrictions $\sum_{j=1}^{p}|\beta_j| < s_1$ and $\sum_{j=2}^{p}|\beta_j - \beta_{j-1}| < s_2$. Thus, the method performs variable selection and smooths consecutive β estimates by shrinking the difference. When using the squared penalty on the difference $\sum_{j=2}^{p}(\beta_j - \beta_{j-1})^2 < s_2$, each difference is shrunken, but not put exactly equal to zero. A requirement for this method is that the covariates can be ordered in some meaningful way. The result is a selection of sequences of non-zero coefficients. Li and Li (2008) and Sun and Wang (2012) suggested a network-constraint by shrinking the estimates of β belonging to the same network. Let $X_{i,j}$ be a 3D contact from genomic element i to genomic elements j. The strongest correlated neighbouring contacts would then be $X_{i-1,j}, X_{i+1,j}, X_{i,j-1}, X_{i,j+1}$. With modification to the fused lasso penalty, the difference between $\beta_{i,j}$ for covariate $X_{i,j}$ against each of the four other neighboring coefficient could be penalized. A selection of sequences of 3D contacts differing between the binary outcomes would then be selected. Further, Sun and Wang (2013) presented a R package including penalized logistic regression, where all neighboring pairs of β estimates can be included as an argument in the R function.

In Jin et al. (2013), the mentioned Hi-C data was analyzed by combining the Hi-C measurements before and after treatment, and pooling all 6 replicates together. Next, they grouped all 3D contacts with similar expectation, distance and properties such as mapping ability and GC content. For a group g they fitted a negative binomial distribution $NB_g(X|\mu_g, \theta_g)$, which were used to calculate the significance of each observed contact frequency by $p_{ij} = NB_g(X > X_{ij}|\mu_g, \theta_g)$ for all X_{ij} belonging to group g. The p-values were corrected for multiple testing by controlling the false discovery rate.

On the other hand, if penalized regression was used to analyze the Hi-C data, one performs a multivariate analysis using all covariates simultaneously, includes all replicates individually, and incorporates the correlation structure in the model. To our knowledge, this theory has not been used on Hi-C data. In such an analysis, the number of covariates (the 3D contacts between each involved genomic element) could be enormous, when the number of genomic elements grows. This may demand a large number of independent samples, larger that the data

set presented here.

7 Discussion

Paper 1 presented a permutation test and checked, using simulations, that the test truly has significance level α. In addition, it is valuable to check the methods ability to find the true negative, thus correctly rejecting the null hypothesis. It is not obvious how to do this in a permutation setting, so this topic will be discussed in this section. Next, we take a closed look at the similarities and differences between lasso and ridge regression. Lastly in this discussion section, important aspects in integration methods are being evaluated.

Challenges when evaluating permutation tests

As mentioned in Section 4.1.1, it is of interest to quantify the power of the permutation test, i.e. the proportion of correctly rejected null hypothesis. This is challenging when the distribution under the alternative hypothesis H_A is unknown. One could try to simulate from an approximation of the distribution under the alternative hypothesis. But, it can be challenging to find a proper simulation setting approximating the real data setting. For the test in Paper I, there is no obvious candidate for a simulation study since the distribution under H_A is highly complex and differs for each observed query set, for different true co-localization signal and for each biological data set. A simplified simulation setting could use modifications of a random walk where the walk has a tendency to move back towards a central location (Gillespie, 1996), or a self-attracting 3-dimensional random walk (Bolthausen and Schmock, 1997). But this would be a highly simplified version of a real biological data set and would not necessarily reflect the true power of our test when testing on real data.

An alternative approach done by Witten and Noble (2012), is to use real biological data, find the most extreme observations and check that those observations actually generate a low and significant p-value. But, note that by the construction of a permutation test the 5% largest (most extreme) observations will always result in the correct conclusion, namely rejection of the hypothesis.

The basic concept of statistical power is that the test should reject H_0 when H_A is true, see Equation (4.2). Therefore, it is a good idea to validate known results where co-localization is present. The hypothesis test in paper I shows, as expected, that regions marked by “active promoter” or “strong enhancer” are significantly co-localized, even after taking the domain properties of the query set into account. Also, when shifting the elements in the query set away from their original positions, the significance is lost and lower enrichment scores are achieved. The test also finds that query sets in either end of the chromosome arms have larger 3D co-localization than genomic elements in the middle of the chromosome arms, as mentioned in Imakaev et al. (2012).

Generally, it is most natural to calculate the power of a test if there are competing methods. To our knowledge, at the time paper I was published, comparable methods were not available.

Lasso versus ridge regression

Lasso and ridge regression are often presented as competing methods and some researchers tend to prefer one. But, what are the main differences, and when should one method be preferred over the other? Table 7.1 gives a brief overview of the differences to be discussed in this section. Firstly, prediction error is standard to compare the performance of the methods. Uniformly, no method dominates the other in terms of prediction (Tibshirani, 1996; Fu, 1998), but there are some special cases worth looking into. In the classical situation with few variables $n > p$ and highly correlated predictors, ridge tends to predict better than lasso (Tibshirani, 1996). The same is the case for $n < p$ and little or no sparsity (Hastie et al., 2009), for instance with gene expressions data (Bøvelstad et al., 2007; van Wieringen et al., 2009). One reason for this could be that ridge regression includes all variables, while lasso works as a variable selection method. When lasso is compared to other variable selection methods, it performs well (Tibshirani, 1996; Hastie et al., 2009). For larger number of covariates ($p >> n$), with or without sparsity, both methods seem to have problems (Zou and Hastie, 2005). Therefore, the "bet on sparsity" principle is formulated by (Hastie et al., 2009) "Use a procedure that does well in sparse problems, since no procedure does well in dense problems". All in all, there is no overall optimal method in terms of prediction.

An important aspect mentioned earlier, is that lasso assumes the true model to be sparse, and thus performs variable selection. If one is unsure whether the sparsity assumption is valid or not, it can be an idea to run the ridge regression first and evaluate the size of the resulting estimates. But in some situations without true sparsity, variable selection is highly desirable and appealing for interpretational reasons (Zou and Hastie, 2005). It is important to have in mind, that the lasso can be unstable for different folds in the K-fold CV and will give different selected covariates (Ein-Dor et al., 2005; Michiels et al., 2005; Meinshausen and Bühlmann, 2010). These different sets may give similar prediction values, but for interpretational reasons it is important to not over interpret the obtained list of selected covariates.

The value of $\hat{\boldsymbol{\beta}}$ depends on the collinearity among the predictors. Ridge regression smooths correlated variables, and in the special case of k identical predictors, each predictor gets identical coefficients with $1/k$ the value that any single predictor would get if fitted alone (Friedman et al., 2010). Lasso tends to select only one variable out of a set of highly correlated covariates (Zou and Hastie, 2005). Both methods shrink the estimates, and for orthonormal $\boldsymbol{X}$, ridge regression shrinks on a multiplicative scaling and lasso on an additive scale.

Lastly, it is worth mentioning again, that version of the lasso possesses the oracle property that the method can correctly select the nonzero coefficients with probability converging to one. Note that some of the assumptions are hard to achieve, for instance that the number of true nonzero coefficients is less than n (Bühlmann and van de Geer, 2011).

Zou and Hastie (2005) suggested the elastic net, having penalty $\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$. It smooths the coefficients, is more stable and still performs variable selection, where correlated variables have a tendency to be either all inn or all out. This may be a good option if one wants to strike a balance between lasso and ridge regression.

Characteristics	Lasso	Ridge
Prediction	Depends on data	Depends on data
Variable selection (sparsity assumtion)	Yes	No
Stability (in K-fold CV)	Unstable	Less unstable
Collinearity	Tends to select one representative	Smooth correlated variables
Shrinkage (orthonormal X)	additive : $\text{sign}(\hat{\beta}_j)(\hat{\beta}_j - \lambda)^+$	multiplicative : $\frac{1}{1+\lambda}\hat{\beta}$
Oracle property	Some versions	No

Table 7.1: The differences between lasso and ridge regression.

Requirement for integration analysis

Data integration has been an increasingly popular field of research the last years, and a number of methods have been proposed. In a regression setting, one option could be to include all variables from all data sets as covariates in the integrated regression. A highly reasonable requirement for integration methods, where one has some sort of co-data, should be to evaluate whether or not to include the extra information. In some situations, the co-data contributes to additional understanding and a stronger signal, but in other cases the co-data only adds more noise to the analysis. For the two methods proposed in paper III and IV, a tuning parameter controls whether or not the co-data is included. This is a strength. One should not automatically assume that integration is better than analyzing the data sets separately.

The two methods, adaptive group-regularized ridge regression and the weighted lasso with data integration, diverge in their perspective on shared signal between the main data source and the co-data. When using the group-regularized ridge, the original β estimates from classical ridge regression need to show different and distinct signals for at least two of the groups defined by the co-data. In other words, if we believe that there are 4 groups of covariates with different impact on the response, the β estimates from the classical ridge, should reflect, to some extent, these differences also without using the group information. If that is the case, we are more convinced that the group information based on the co-data is true. The weighted lasso with data integration is not dependent on the signal in the classical lasso with no data integration. In situations when the classical lasso does not find any signal in the main data set, the co-data could still be incorporated and give overall better predictions. This can be preferable when there is too much noise in the main data, but a clear signal based on the co-data.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188.

Bergersen, L. C., Glad, I. K., and Lyng, H. (2011). Weighted lasso with data integration. *Stat. Appl. Genet. Mol. Biol.*, 10(1):1–29.

Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E. W., Wu, B., Doucet, D., Thomas, N. J., Wang, Y., Vollmer, E., et al. (2006). High-throughput dna methylation profiling using universal bead arrays. *Genome research*, 16(3):383–393.

Bird, A. (2002). Dna methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21.

Bolthausen, E. and Schmock, U. (1997). On self-attracting d-dimensional random walks. *Ann. Probab.*, 25(2):531–572.

Botta, M., Haider, S., Leung, I. X., Lio, P., and Mozziconacci, J. (2010). Intra-and inter-chromosomal interactions correlate with ctcf binding genome wide. *Molecular systems biology*, 6(1):426.

Bøvelstad, H., Nygård, S., Størvold, H., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjærde, O. (2007). Predicting survival from microarray data - a comparative study. *Bioinformatics*, 23(16):2080–2087.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer-Verlag, Heidelberg. Methods, theory and applications.

Calladine, C. R., Drew, H., Luisi, B., and Travers, A. (2004). *Understanding DNA: the molecule and how it works*. Academic Press.

Cheah, P. and Looi, L. (2001). p53: an overview of over two decades of study. *The Malaysian journal of pathology*, 23(1):9–16.

Collins, F. S., Morgan, M., and Patrinos, A. (2003). The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290.

Cule, E., Vineis, P., and De Iorio, M. (2011). Significance testing in ridge regression for genetic data. *BMC bioinformatics*, 12(1):372.

Dai, Z. and Dai, X. (2010). Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast. *Nucleic Acids Res.*, 40:27–36.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science*, 295:1306–1311.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485:376–380.

Dobson, A. J. and Barnett, A. (2008). *An Introduction to Generalized Linear Models, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis Group, third edition.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., and et. al. (2006). Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16:1299–1309.

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A., and Noble, W. S. (2010). A three-dimensional model of the yeast genome. *Nature*, 465:363–367.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2):407–499. With discussion, and a rejoinder by the authors.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *J. Amer. Statist. Assoc.*, 68:117–130.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178.

Erhuma, A. M. (2012). Glucocorticoids: Biochemical group that play key role in fetal programming of adult disease. `http://cdn.intechopen.com/pdfs-wm/41149.pdf`. [Online; accessed 11-June-2015].

Francis, R. (2011). *Epigenetics : the ultimate mystery of inheritance*. W. W. Norton & Company, Inc.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist.*, 7(3):397–416.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., and Mei, P. H. e. a. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462:58–64.

Garcia, P. T., Müller, S., Carroll, R. J., Dunn, T. N., Thomas, A. P., Adams, A. H., Pillai, S. D., and Walzem, R. L. (2013). Structured variable selection with q-values. *Biostatistics*, 14(4):695–707.

Gillespie, D. T. (1996). Exact numerical simulation of the ornstein-uhlenbeck process and its integral. *Phys. Rev. E*, 54:2084–2091.

Goeman, J. J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biom. J.*, 52(1):70–84.

Goeman, J. J., van de Geer, S. A., and van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):477–493.

Goldberg, A. D., Allis, C. D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, 128(4):635–638.

Good, P. (2005). *Permutation, parametric and bootstrap tests of hypotheses*. Springer Series in Statistics. Springer-Verlag, New York, third edition.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.*, 18(17-18):2529–2545.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, second edition. Data mining, inference, and prediction.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, pages 337–344.

Hoerl, A. E. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., and Mirny, L. A. (2012). Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature*, 9:999–1003.

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C.-A., Schmitt, A. D., Espinoza, C. A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–294.

Kleinjan, D. and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *The American Journal of Human Genetics*, 76:8–32.

Le Cessie, S. and Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 41(1):191–201.

Lehmann, E. L. (1999). *Elements of large-sample theory*. Springer Texts in Statistics. Springer-Verlag, New York.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for dna methylation in genomic imprinting. *Nature*, 366:362–365.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., and Dorschner, M. O. e. a. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.

Margueron, R. and Reinberg, D. (2010). Chromatin structure and the inheritance of epigenetic information. *Nature Reviews Genetics*, 11(4):285–296.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London. Second edition [of MR0727836].

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(4):417–473.

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, 365(9458):488–492.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *J. Amer. Statist. Assoc.*, 78(381):47–65.

Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008). *P*-values are random variables. *Amer. Statist.*, 62(3):242–245.

Nature Education (2010). Nature education, scitable: Gene expression. `http://www.nature.com/scitable/topicpage/gene-expression-14121669`. [Online; accessed 11-June-2015].

Paulsen, J., Sandve, G. K., Gundersen, S., Lien, T. G., Trengereid, K., and Hovig, E. (2014). HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. *Bioinformatics*, 30(11):1620–1622.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., and Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American journal of epidemiology*, 159(9):882–890.

Pertea, M. and Salzberg, S. L. (2010). Between a chicken and a grape: estimating the number of human genes. *Genome Biol.*, 11:206.

Phipson, B. and Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.*, 9(1).

Ptashne, M. (2007). On the use of the word 'epigenetic'. *Current Biology*, 17(7):R233–R236.

Quina, A., Buschbeck, M., and Di Croce, L. (2006). Chromatin structure and epigenetics. *biochemical pharmacology*, 72(11):1563–1569.

Reppe, S., Refvem, H., Gautvik, V. T., Olstad, O. K., Høvring, P. I., Reinholt, F. P., Holden, M., Frigessi, A., Jemtland, R., and Gautvik, K. M. (2010). Eight genes are highly associated with bmd variation in postmenopausal caucasian women. *Bone*, 46(3):604–612.

Schones, D. E. and Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics*, 9(3):179–191.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1–13.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.*, 22(2):231–245.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nature Genetics*, 38:1348–1354.

Smith, C. M., Haimberger, Z. W., Johnson, C. O., Wolf, A. J., Gafken, P. R., Zhang, Z., Parthun, M. R., and Gottschling, D. E. (2002). Heritable chromatin structure: mapping "memory" in histones h3 and h4. *Proceedings of the National Academy of Sciences*, 99(suppl 4):16454–16461.

Sun, H. and Wang, S. (2012). Penalized logistic regression for high-dimensional dna methylation data with case-control studies. *Bioinformatics*, 28(10):1368–1375.

Sun, H. and Wang, S. (2013). Network-based regularization for matched case-control analysis of high-dimensional dna methylation data. *Stat. Med.*, 32(12):2127–2139.

Tai, F. and Pan, W. (2007). Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. *Bioinformatics*, 23(14):1775–1782.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J., and Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Stat. Med.*, 25(18):3201–3216.

van Wieringen, W. N., Kun, D., Hampel, R., and Boulesteix, A.-L. (2009). Survival prediction using gene expression data: a review and comparison. *Comput. Statist. Data Anal.*, 53(5):1590–1603.

Wagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M. (2014). The relationship between dna methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol*, 15(2):R37.

Watson, J. D. (1990). The human genome project: past, present, and future. *Science*, 248(4951):44–49.

West, A. and Fraser, P. (2005). Remote control of gene transcription. *Human Molecular Genetics*, 14:R101–R111.

Witten, D. and Noble, W. (2012). On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Res.*, 40:3849–3855.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67.

Ziegler, A., König, I. R., and Pahlke, F. (2010). *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an e-learning platform*. John Wiley & Sons.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320.